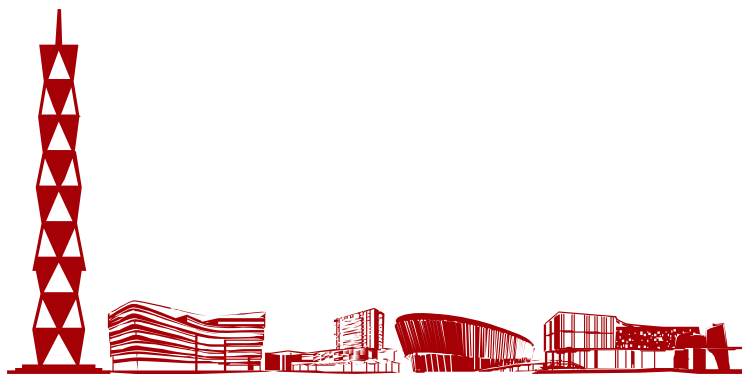


PMEM Throughput Aggregation with Non-volatile Page Table Tree

Xiaolong Guo, Qinzhou Zhang and Shu Yin
ShanghaiTech University



Contents



上海科技大学
ShanghaiTech University

1 Background and Motivation

2 Key Idea

3 Design

4 Evaluation

5 Discussion

6 Conclusion and Future Work



立志成才 报国裕民

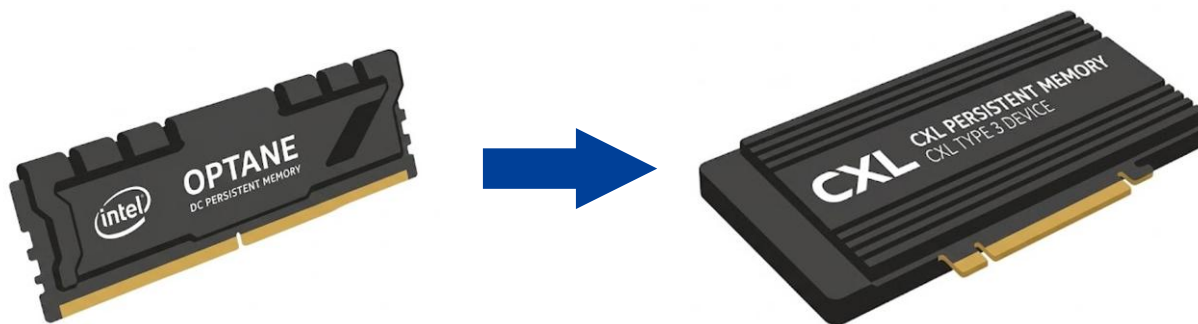
Background & Motivation



上海科技大学
ShanghaiTech University

Attractive properties of PMEM:

- Persistent
- Low Latency
- Byte Addressable



CXL-PMEM is expected to be the next generation of PMEM.

CXL-PMEM can provide scalability, support hot plug.

Background & Motivation



上海科技大学
ShanghaiTech University

PMEM is a multi-device technology.

A single PMEM module offers limited bandwidth and may suffer from contention under concurrent access.

Typically, we utilize multiple PMEM devices organized in an interleaved manner.

- **Multiple PMEMs deliver higher bandwidth.**
- **Multiple PMEMs can reduce contention.**



立志成才 报国裕民

Background & Motivation

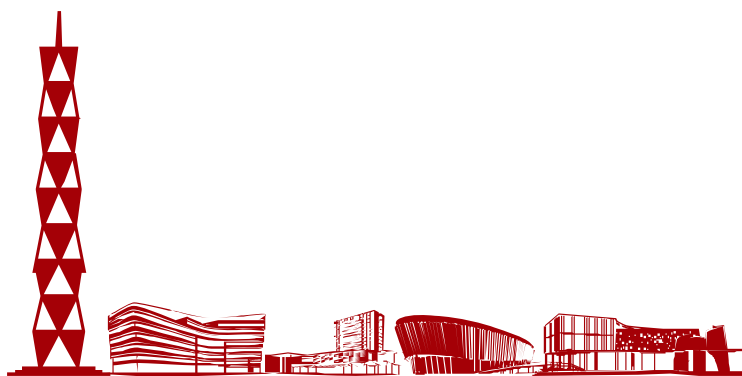


上海科技大学
ShanghaiTech University

Hardware Interleave is not flexible:

Cannot span NUMA nodes.
Cannot be reconfigured at runtime.

Cannot adapt to the CXL-PMEM era.



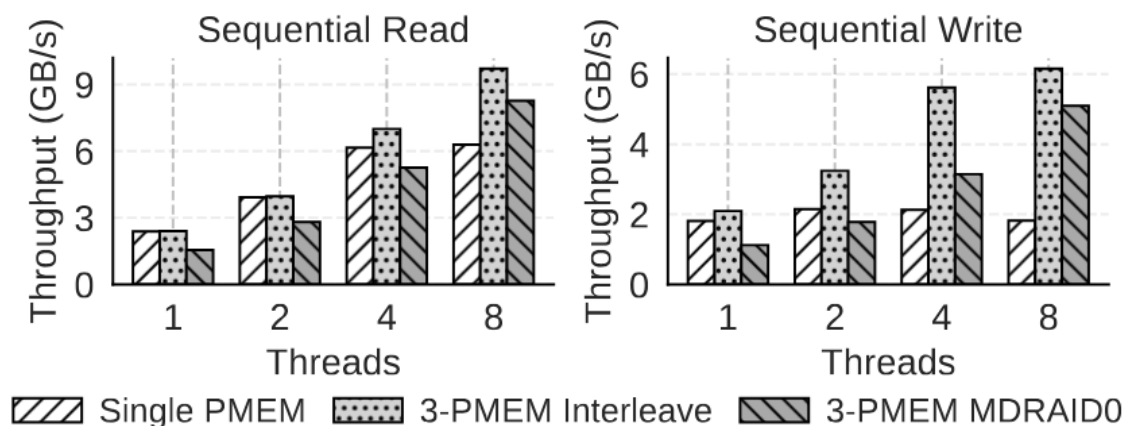
立志成才 报国裕民

Background & Motivation



上海科技大学
ShanghaiTech University

Two different methods for utilizing the throughput of multiple PMEM modules:
Hardware interleave and **mdadm RAID0**



- Hardware interleaving multiple PMEMs improves performance.
- Software RAID 0 falls far behind hardware interleaving

Ext4 on Different PMEM Configurations with 4KB IO Size.



立志成才 报国裕民

Background & Motivation



上海科技大学
ShanghaiTech University

Existing Multi PMEM Throughput Aggregation Methods: A Trade-off

Hardware Interleaving

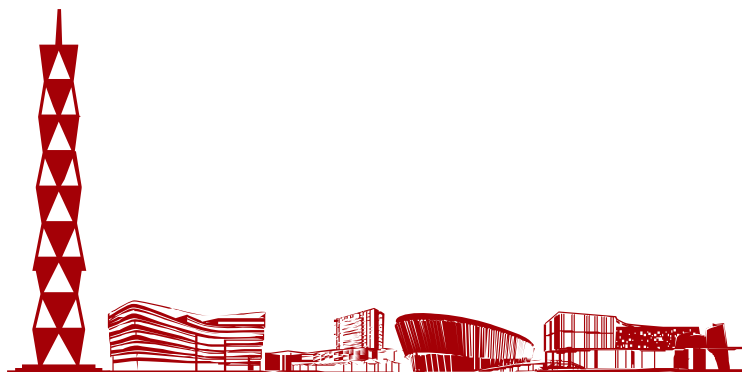
- **High performance**
but can not support:
- **Dynamic configuration.**
- **Cross NUMA nodes.**

Efficient but not flexible.

Software RAID0

- **High flexibility**
but suffer from:
- **Low performance.**
- **DAX-MMAP missing.**

Flexible but not efficient.



立志成才 报国裕民

Background & Motivation

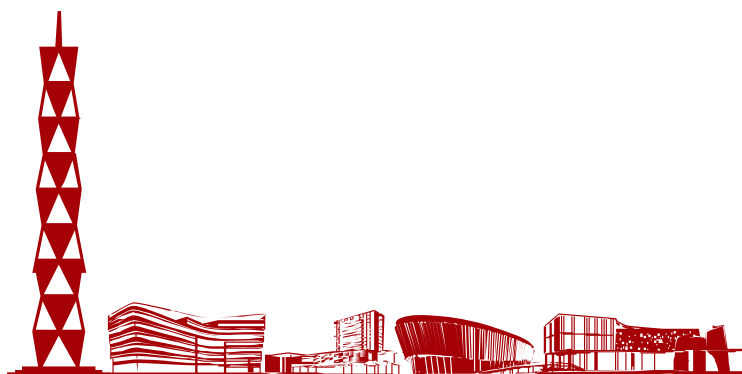


上海科技大学
ShanghaiTech University

Existing Multi PMEM Throughput Aggregation Methods: An Unfavorable Trade-off

Problem:

How to efficiently and flexibly manage multiple PMEM devices?



立志成才 报国裕民

Key Idea



上海科技大学
ShanghaiTech University

- Virtual addresses to physical addresses translation using page table.
- Page table is efficient because of the MMU and TLB support.
- Page table mappings can be dynamically adjusted.



**Memory management may have solution.
Page table can be data structure.**

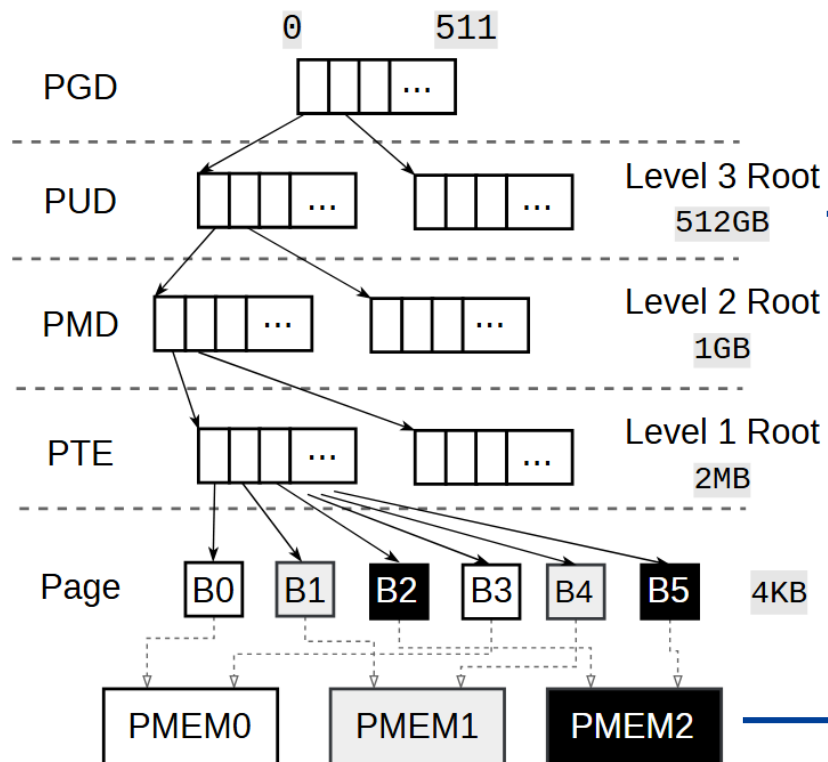


立志成才 报国裕民

Key Idea



上海科技大学
ShanghaiTech University



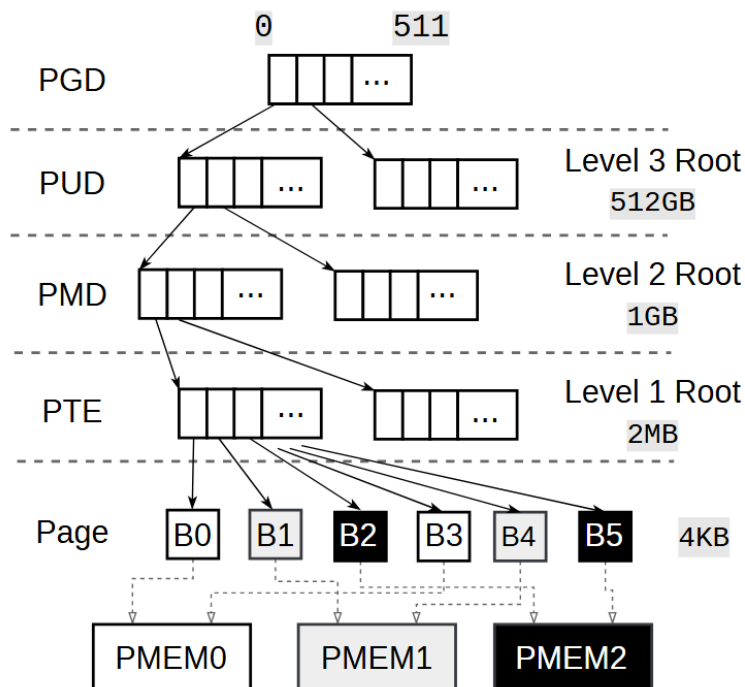
A radix tree which has exactly the same structure as normal page table, but it is persistent. Can be installed to virtual address of a process.

Data managed by NVPTT are interleaved across multi PMEMs.

NVPTT: A data structure that can be accelerated by MMU and TLB!!!

立志成才 报国裕民

Vessel NVPTT



Multiple Level Mechanism

NVPTT has different levels, each representing an addressing range of a different size.

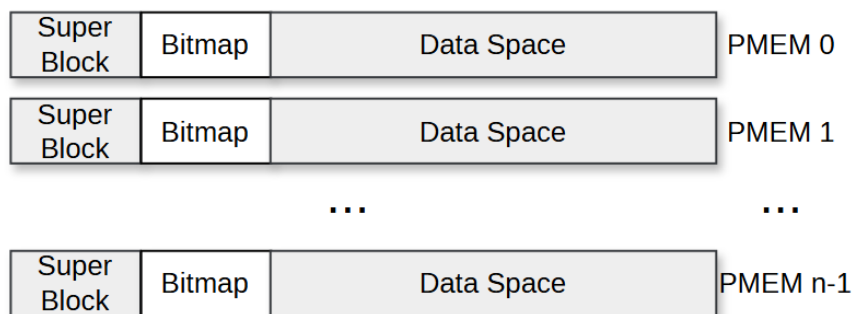
VMA based Virtual Space Management

Each NVPTT installed into a process corresponds to a VMA, leveraging Linux's existing virtual address space management mechanism.

Page Fault based Append

NVPTT allocates space lazily on demand via page faults, allocating 2MB at a time.

Vessel Space Management

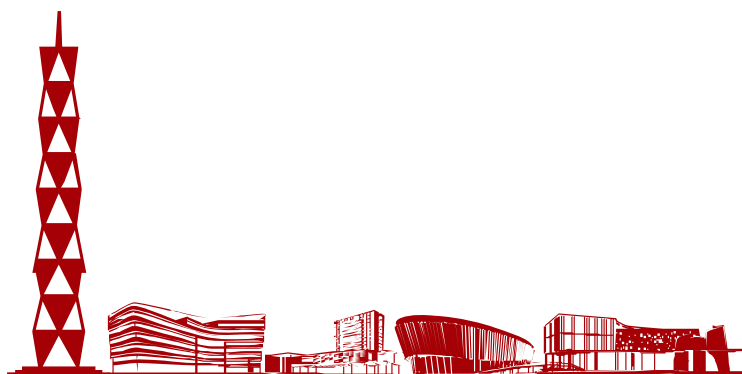


Symmetric Space Layout.

The multiple PMEM devices managed by Vessel share the same space layout, which includes a superblock area, a bitmap area, and a data space area.

Block Allocator.

Vessel has two different block allocators: the Round-Robin Block Allocator and the Interleaved Block Allocator.



Vessel Programming Model

APIs is exposed through ioctl of VFS.

vessel_add_pmem: Add PMEM to Vessel.

vessel_new_ptt: Create a new nvptt.

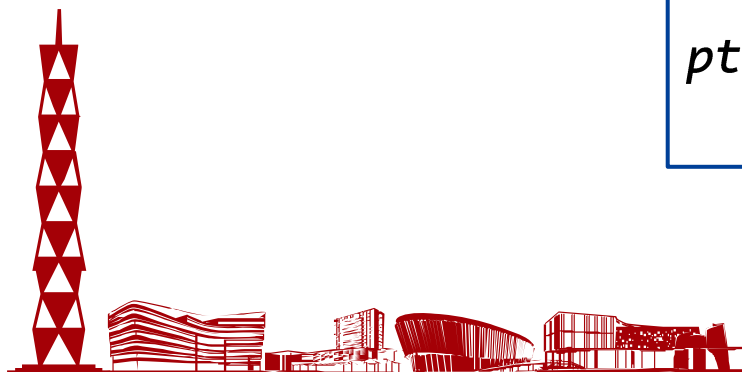
vessel_ptt_map: Map a NVPTT to process.

vessel_ptt_unmap: Unmap a NVPTT.

vessel_del_ptt: Delete a NVPTT.

ptt_rebalance: Rebalance a NPTTT.

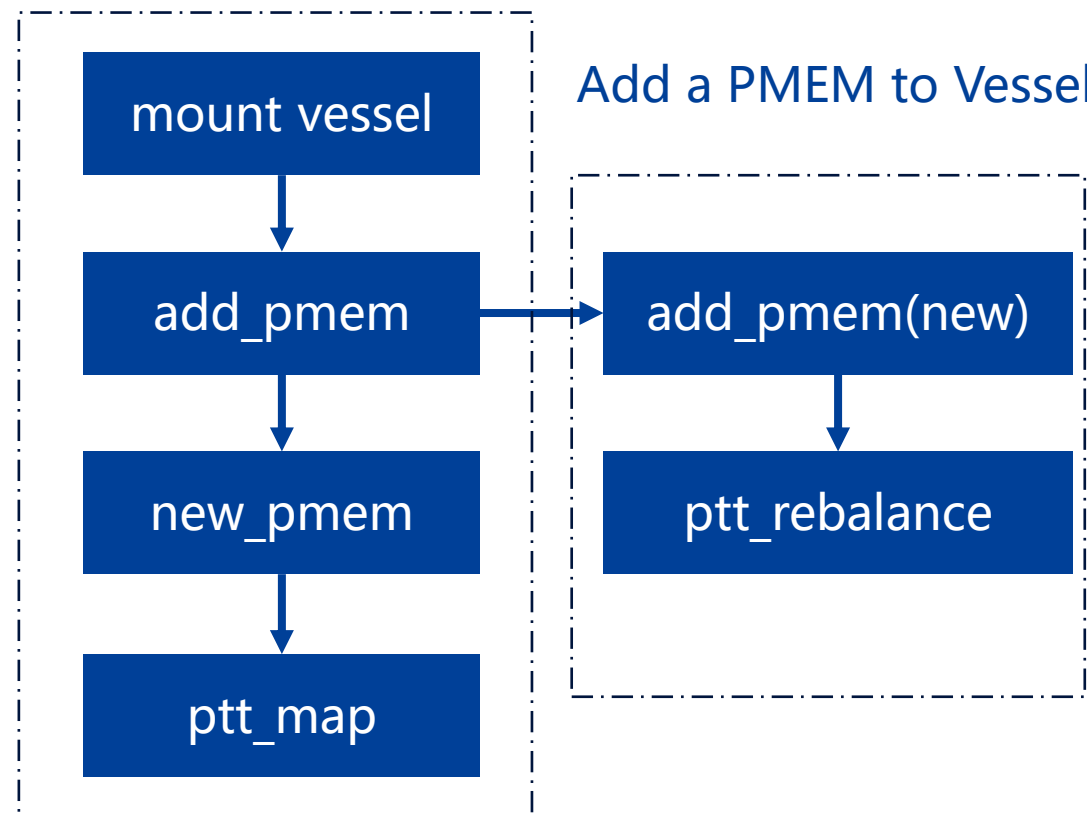
ptt_level_up: Level up a NVPTT.



Vessel Programming Model

vessel_add_pmem: Add PMEM to Vessel.
vessel_new_ptt: Create a new nvptt.
vessel_ptt_map: Map a NVPTT to process.
vessel_ptt_unmap: Unmap a NVPTT.
vessel_del_ptt: Delete a NVPTT.
ptt_rebalance: Rebalance a NPTTT.
ptt_level_up: Level up a NVPTT.

Map an NVPTT to process



Evaluation Setup



上海科技大学
ShanghaiTech University

Hardware:

Intel Xeon Gold 6240M CPU (18 cores)

256G Intel Optane PMEM 100 Series * 3

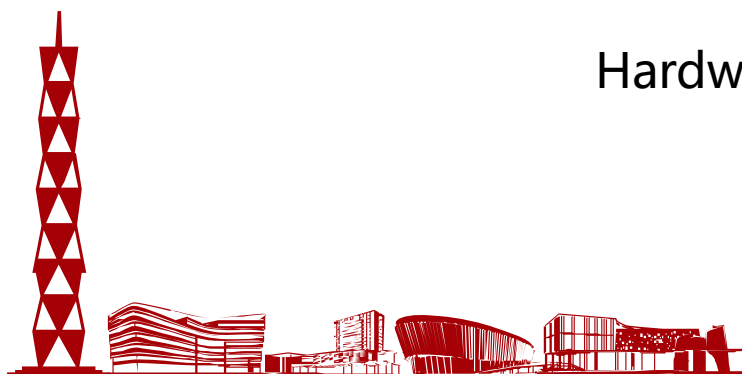
Operating System :

Ubuntu 24.04 with linux 6.4.0

Compare With:

Software RAID0 (MADAM RAID0 + EXT4)

Hardware Interleave (Interleaved 3 PMEMs, devdax mode, mmap)



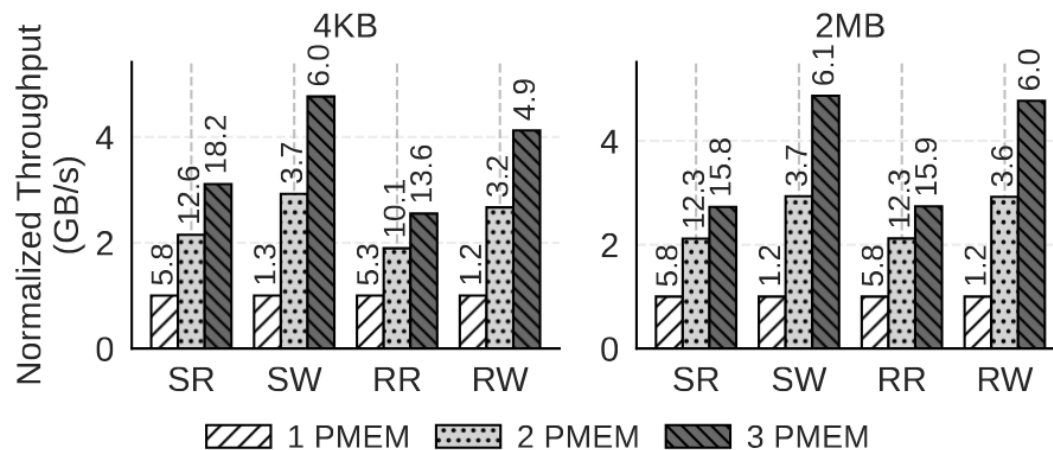
立志成才 报国裕民

Evaluation



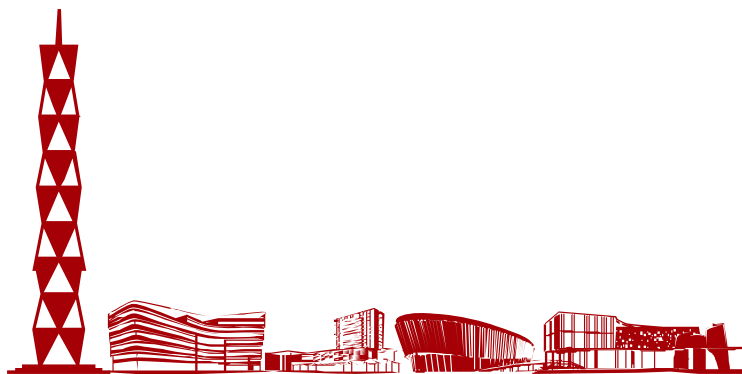
上海科技大学
ShanghaiTech University

Performance of Different Numbers of PMEMs



More PMEM devices can deliver better performance.

Vessel with Different Number of PMEMs at 8 Threads



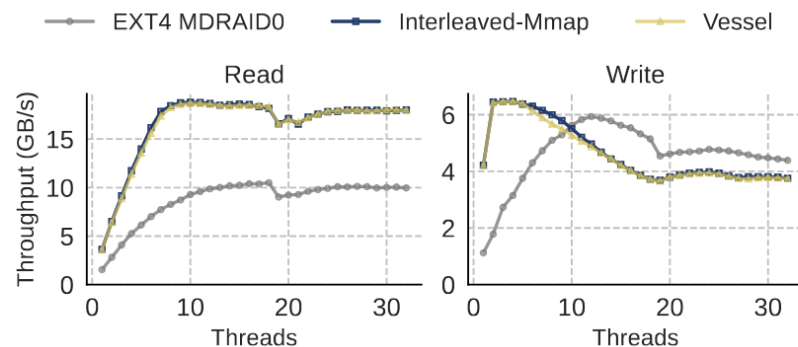
立志成才 报国裕民

Evaluation

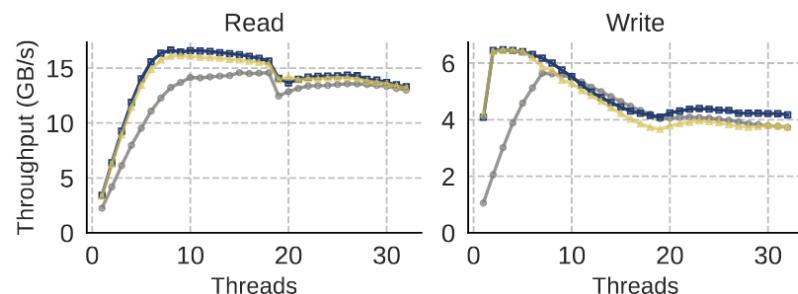


上海科技大学
ShanghaiTech University

Throughput vs IO Size and #Thread



(a) Results with 4KB IO size



(b) Results with 2MB IO size

O1: With a small number of threads, interleaved mmap significantly outperforms EXT4 on mdadm RAID0

O2: As thread count grows, at 4KB I/O, EXT4 trails interleaved mmap in reads but leads in writes, at 2MB I/O, the gap between systems narrows.

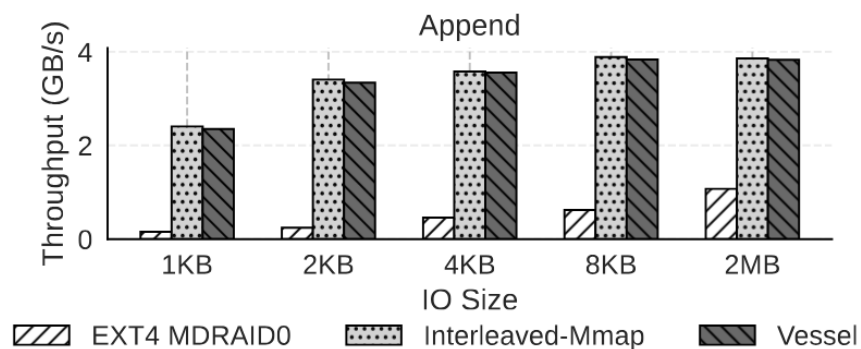
O3: Vessel consistently achieves performance close to that of interleaved mmap, even though its page table itself resides in PMEM.

Evaluation

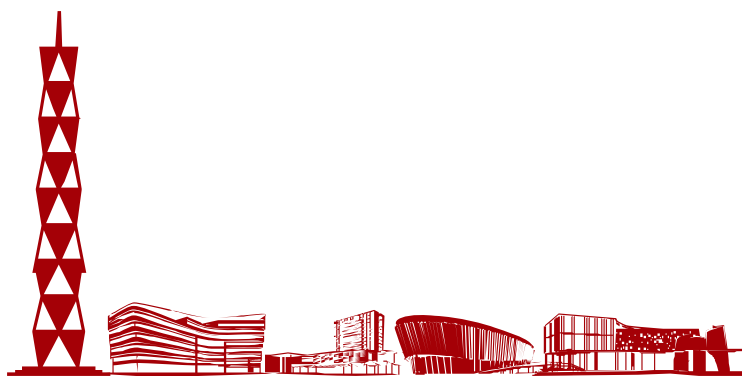


上海科技大学
ShanghaiTech University

Performance of Append



Vessel's append performance is also close to that of Interleaved MMap and far exceeds that of EXT4 built on mdadm RAID0.



立志成才 报国裕民

Evaluation



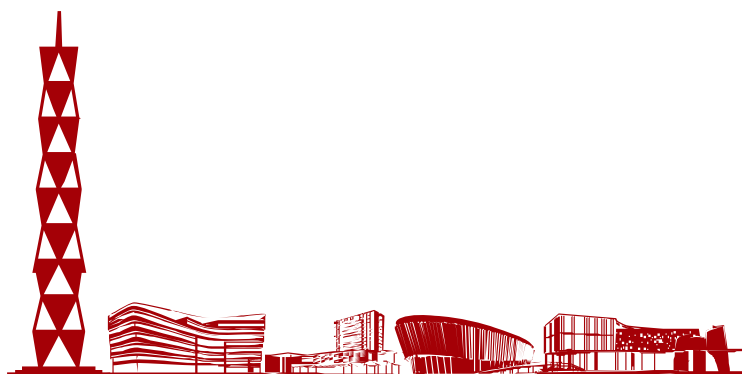
上海科技大学
ShanghaiTech University

Time to rebalance after adding PMEMs(s)

NVPTT Data Size	1 PMEM → 2 PMEMs	1 PMEM → 3 PMEMs
256MB	0.20	0.18
1GB	0.95	0.66
4GB	3.74	2.58
16GB	14.95	10.23

Vessel supports adding devices without shutting down.

After rebalance, NVPTT can distribute data across the newly added devices. The larger the data in NVPTT, the longer the rebalancing takes.



立志成才 报国裕民

Discussion



上海科技大学
ShanghaiTech University

CXL Compatibility.

Vessel' s design relies only on byte-addressability and persistence, It is applicable to CXL-attached PMEM.

Comparison with CXL Hardware Interleaving.

Reprogramming the hardware decoders could adjust CXL memory interleaving but it requires a full data migration. NVPTT based interleaving does not.

Stripe Granularity.

The stripe size of NVPTT can only be 4KB or a multiple of it, but this is a very practical I/O size.



立志成才 报国裕民

Conclusion & Future Work



上海科技大学
ShanghaiTech University

Conclusion

We present Vessel, a PMEM pooling and mapping system for flexible multi-PMEM management and throughput aggregation.

Vessel shows that interleaving pmem at the virtual address level can achieve performance close to physical-address interleaving, while offering better flexibility.

Future Work

Vessel will serve as the substrate for our future PMEM file system, aimed at managing multiple PMEM devices.



立志成才 报国裕民



The 18th ACM Workshop on
Hot Topics in Storage and File Systems

Sept. 28-29
Prague, Czechia
Co-located with SOSP 2026

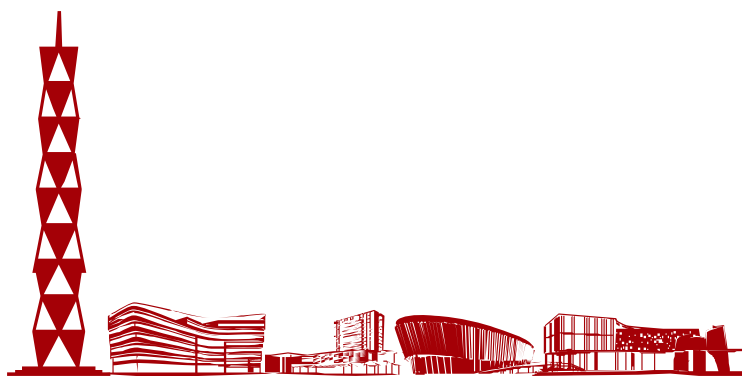


L.I.O.N
LABORATORY OF
I/O SYSTEMS
& DATA SCIENCE



上海科技大学
ShanghaiTech University

Thank You!



立志成才 报国裕民